$a \cdot (x \cdot x)$ or $(a \cdot x) \cdot x$?

Jean-Michel Muller

$a \cdot (x \cdot x)$ or $(a \cdot x) \cdot x$?

Jean-Michel Muller
CNRS, LIP, Université de Lyon
Lyon, France
jean-michel.muller@ens-lyon.fr

***Abstract*—Expressions such as ax^2, axy, or ax^3, where a is a constant, are not unfrequent in computing. There are several ways of parenthesizing them (and therefore, choosing the order of evaluation). Depending on the value of a, is there a more accurate evaluation order? We discuss this point (with a small digression on spurious underflows and overflows).**

***Index Terms*—Floating-point arithmetic, rounding error analysis, evaluation of expressions.**

Introduction and notation

In this paper, which aims more at "floating-point pedagogy" than at complex research, we discuss the evaluation of expressions of the form ax^2, axy, or ax^3 in binary floating-point arithmetic, where a is a constant and x is a variable. There are several ways of parenthesizing these expressions, which correspond to different evaluation orders (for instance, ax^2 may be evaluated as $(a \cdot x) \cdot x$—i.e., ax is evaluated first—or as $a \cdot (x \cdot x)$—i.e., x^2 is evaluated first). A question that naturally arises is: *is one of these schemes better than the other ones?* The word "better" may have different meanings here, because several criterions are possible: one may wish to maximize parallelism, in the hope of having a faster evaluation;[1] one may try to minimize the relative error of the result, in the hope of making the whole calculation more accurate; or one may try to avoid as much as possible spurious underflows or overflows (a *spurious underflow or overflow* is an underflow or overflow that occurs during an intermediate step, resulting in an inaccurate, infinite or NaN returned result, whereas the exact result is well within the domain of normal floating-point numbers). Deciding the adequate order of evaluation will be in general the programer's task (through unambiguous parenthesizing and/or adequate compilation options). In some cases (no parenthesizing and the language specification does not impose in that case an order of evaluation, which is for instance the case of FORTRAN [2], [5]) it can be the compiler's task and one of the goals of this paper is to persuade compiler designers that is such cases, attempting to minimize the evaluation delay is not always the only sensible option at hand.

In the following, we assume a radix-2, precision-p, floating-point (FP) arithmetic [1], [5]. The notation $\mathrm{RN}(t)$ stands for t rounded to the nearest FP number. We assume no particular tie-breaking rule in our proofs, and our examples are generated assuming ties-to-even. The number $u = 2^{-p} = \frac{1}{2}\mathrm{ulp}(1)$ denotes the roundoff error unit. If x is a nonzero real number, with $2^k \leq |x| < 2^{k+1}$ (i.e., 2^k is what Rump defines as $\mathrm{ufp}(x)$ [6]), $\mu(x) = x/2^k$ denotes the "infinite-precision significand" of x. Barring underflow or overflow, the relative error due to rounding to nearest a nonzero real number x, i.e.,

$$\left|\frac{\mathrm{RN}(x) - x}{x}\right|$$

is bounded by $u/\mu(x)$. This allows one for instance to show that the relative error due to rounding is bounded by $u/(1+u)$ [4], and to show very tight bounds on the relative errors of Floating-Point operations [3], [6]. We assume that all intermediate calculations are performed in the same format.

Take as an example the computation of ax^2 as $a \cdot (x \cdot x)$. What is actually computed is $\mathrm{RN}(a \cdot \mathrm{RN}(x \cdot x))$. If we denote ϵ_s the relative error of the square and ϵ_m the relative error of the subsequent multiplication, the computed result r satisfies

$$r = ax^2 \cdot (1+\epsilon_s)(1+\epsilon_m) = ax^2 \cdot (1 + \epsilon_s + \epsilon_m + \epsilon_s\epsilon_m).$$

In the following, we will consider *order-1* errors. That is, we will approximate the above-given relative error $\epsilon_s + \epsilon_m + \epsilon_s\epsilon_m$ by $\epsilon_s + \epsilon_m$. Order-1 error bounds are not enough if one wishes an *absolute certainty* that the error is less than the bound, but they suffice for comparing computation schemes, which is the aim of this article.

I. Computation of ax^2

For symmetry reasons, we assume that a and x are positive. We also assume that they are (strictly) between 1 and 2. This is done without loss of generality provided that no underflow or overflow occurs in the calculation.[2]

A. First choice: $a \cdot (x \cdot x)$

We start by computing the square of x. The relative error of that operation is bounded by $u/\mu(x^2)$, i.e., by

$$\begin{cases} \frac{u}{x^2} & \text{if} \quad x < \sqrt{2}, \\ \frac{2u}{x^2} & \text{if} \quad x > \sqrt{2}. \end{cases} \tag{1}$$

[1] This is not a trivial issue: in practice, the parenthesizing may impact register allocation and sub-expression sharing.

[2] We define underflow as the IEEE 754 Standard does: an inexact zero or subnormal result.

(what we choose for $x = \sqrt{2}$ does not matter: since x is a floating-point number, if *cannot* be equal to $\sqrt{2}$).

The relative error of the second operation (namely, the multiplication of the previously-obtained square by a) is bounded by $u/\mu(ax^2)$. Since ax^2 is between 1 and 8, the bound $u/\mu(ax^2)$ is equal to

$$\begin{cases} \frac{u}{ax^2} & \text{if} \quad x < \sqrt{\frac{2}{a}}, \\ \frac{2u}{ax^2} & \text{if} \quad \sqrt{\frac{2}{a}} < x < \frac{2}{\sqrt{a}}, \\ \frac{4u}{ax^2} & \text{if} \quad x > \frac{2}{\sqrt{a}}. \end{cases} \tag{2}$$

One easily checks that $\sqrt{\frac{2}{a}} < \sqrt{2}$ and $\sqrt{2} < \frac{2}{\sqrt{a}}$. Hence one can divide the interval $[1, 2)$ where x lies into four subintervals, and in each subinterval compute the total relative error bound (obtained by summing the error bounds of the individual operations, given by (1) and (2)), which is a function of a and x. Then for each subinterval one can compute the maximum value of the relative error bound, which is a function of a only (this is easily done: in each subinterval the bound is a decreasing function of x, so that its maximum is attained at the leftmost point of the interval). This is done in Table I.

TABLE I: Largest relative error in each of the subintervals (we compute $a \cdot (x \cdot x)$).

Interval for x	$\left[1, \sqrt{\frac{2}{a}}\right]$	$\left[\sqrt{\frac{2}{a}}, \sqrt{2}\right]$	$\left[\sqrt{2}, \frac{2}{\sqrt{a}}\right]$	$\left[\frac{2}{\sqrt{a}}, 2\right]$
$\frac{1}{u} \times$ rel. error	$\frac{1}{x^2} + \frac{1}{ax^2}$	$\frac{1}{x^2} + \frac{2}{ax^2}$	$\frac{2}{x^2} + \frac{2}{ax^2}$	$\frac{2}{x^2} + \frac{4}{ax^2}$
$\frac{1}{u} \times$ largest val.	$1 + \frac{1}{a}$	$1 + \frac{a}{2}$	$1 + \frac{1}{a}$	$1 + \frac{a}{2}$

From Table I we immediately deduce that the relative error is bounded (at order 1 in u) by

$$u \times \max\left\{1 + \frac{1}{a}; 1 + \frac{a}{2}\right\} = \begin{cases} (1 + \frac{1}{a}) \cdot u & \text{if} \quad a \le \sqrt{2} \\ (1 + \frac{a}{2}) \cdot u & \text{if} \quad a > \sqrt{2}. \end{cases} \tag{3}$$

Figure 1 plots the bound (3) and the actually obtained largest relative error (for all x) in the case $p = 16$ for $a \in [1, 2)$ (to generate the figure, since for each value of a we need to consider all possible values of x, we are limited to small values of the precision p). The plot shows that the bound (3) is tight.

B. Second choice: $(a \cdot x) \cdot x$

We now start by computing $a \cdot x$. The relative error of that operation is bounded by $u/\mu(ax)$, i.e., by

$$\begin{cases} \frac{u}{ax} & \text{if} \quad x \le \frac{2}{a}, \\ \frac{2u}{ax} & \text{if} \quad x > \frac{2}{a}. \end{cases} \tag{4}$$

The relative error of the second operation is bounded by $u/\mu(ax^2)$, i.e., the same bound as in (2). After having noted that $\sqrt{\frac{2}{a}} < \frac{2}{a}$ and $\frac{2}{a} < \frac{2}{\sqrt{a}}$ we can divide the interval $[1, 2)$ where x lies into four subintervals and compute for each subinterval the maximum of the sum of the bounds (2) and (4). This is done in Table II.

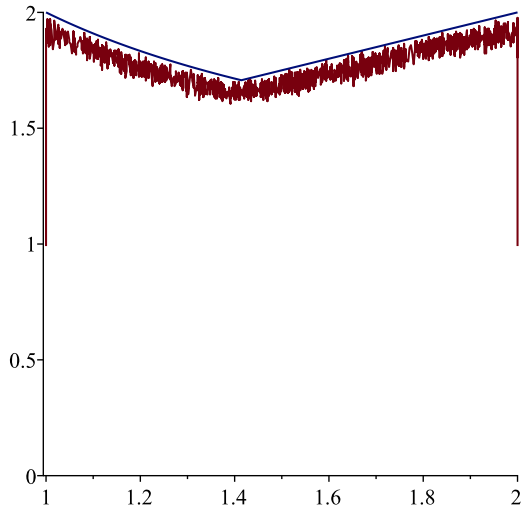

Fig. 1: The relative error bound for the computation of ax^2 as $a \cdot (x \cdot x)$, in multiples of u and as a function of $a \in [1, 2)$, along with the largest actually obtained values for $p = 16$.

TABLE II: Largest relative error in each of the subintervals (we compute $(a \cdot x) \cdot x$).

Interval for x	$\left[1, \sqrt{\frac{2}{a}}\right]$	$\left[\sqrt{\frac{2}{a}}, \frac{2}{a}\right]$	$\left[\frac{2}{a}, \frac{2}{\sqrt{a}}\right]$	$\left[\frac{2}{\sqrt{a}}, 2\right]$
$\frac{1}{u} \times$ rel. error	$\frac{1}{ax} + \frac{1}{ax^2}$	$\frac{1}{ax} + \frac{2}{ax^2}$	$\frac{2}{ax} + \frac{2}{ax^2}$	$\frac{2}{ax} + \frac{4}{ax^2}$
$\frac{1}{u} \times$ largest val.	$\frac{2}{a}$	$1 + \frac{\sqrt{2}}{2\sqrt{a}}$	$1 + \frac{a}{2}$	$1 + \frac{1}{\sqrt{a}}$

Hence the relative error is bounded (at order 1 in u) by

$$u \times \max\left\{\frac{2}{a}; 1 + \frac{\sqrt{2}}{2\sqrt{a}}; 1 + \frac{a}{2}; 1 + \frac{1}{\sqrt{a}}\right\}.$$

Elementary manipulation shows that for $a \in [1, 2)$, $\frac{2}{a} \le 1 + \frac{1}{\sqrt{a}}$ and $1 + \frac{\sqrt{2}}{2\sqrt{a}} \le 1 + \frac{1}{\sqrt{a}}$. Therefore the relative error bound can be simplified and becomes

$$\begin{aligned} & u \times \max\left\{1 + \tfrac{a}{2}; 1 + \tfrac{1}{\sqrt{a}}\right\} \\ &= \begin{cases} \left(1 + \frac{1}{\sqrt{a}}\right) u & \text{if} \quad a \le 2^{\frac{2}{3}} \\ \left(1 + \frac{a}{2}\right) u & \text{if} \quad a > 2^{\frac{2}{3}} \end{cases} \end{aligned} \tag{5}$$

Figure 2 plots the bound (5), and the actually obtained largest relative error (for all x) in the case $p = 16$ for $a \in [1, 2)$.

C. Comparison of the schemes for ax^2

Figure 3 plots the bounds (3) and (5). If $a < 2^{2/3} \approx 1.587$ the bound corresponding to $a \cdot (x \cdot x)$ is smaller, if $a \ge 2^{2/3}$, both bounds are equivalent. This is confirmed

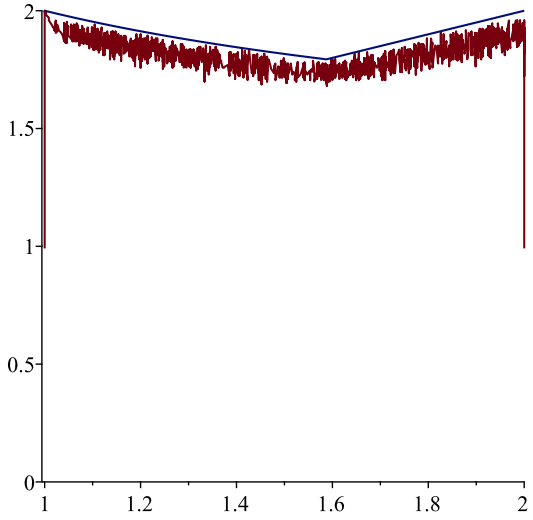

Fig. 2: The relative error bound for the computation of ax^2 as $(a \cdot x) \cdot x$, in multiples of u and as a function of $a \in [1, 2)$, along with the largest actually obtained values for $p = 16$.

by Figure 4, where we have plotted the difference of the actual relative errors of both schemes in the case $p = 16$.

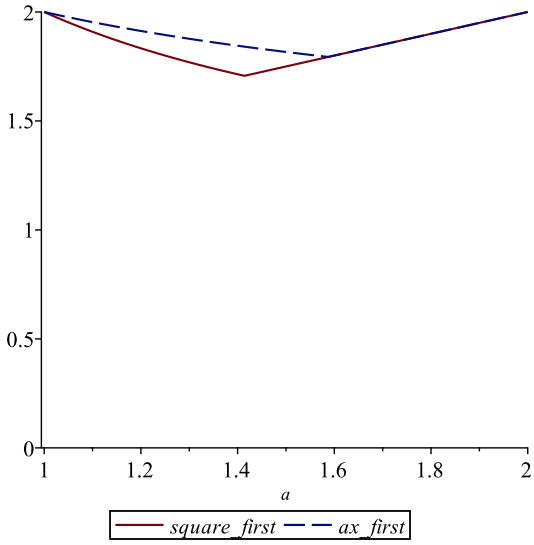

Fig. 3: The two bounds for the computation of ax^2 (multiplied by $1/u$), as a function of $a \in [1, 2)$. If $a < 2^{2/3}$ the bound corresponding to $a \cdot (x \cdot x)$ is smaller than the other one, if $a \geq 2^{2/3}$, both bounds are equal.

Hence we conclude that for a in an interval of the form $2^k \times [1, 2^{2/3}]$, $a \cdot (x \cdot x)$ has a better relative error bound (and, in general, a better actual relative error) than $(a \cdot x) \cdot x$, and that for a in an interval of the form $2^k \times [2^{2/3}, 2)$, the relative error bound is the same (and the actual errors are of similar order) for both schemes.

Does this mean that we should recommend always computing the square first, i.e., using the scheme $a \cdot (x \cdot x)$? This is not that simple. The scheme $(a \cdot x) \cdot x$ has an advantage when it comes to avoiding spurious underflows or overflows. Assume that a is a normal floating-point number. If $a \cdot x$ underflows (i.e., the obtained result is 0 or subnormal) this means that $|x| < 1$, so that $|ax^2|$ is even smaller than $|ax|$. If $a \cdot x$ overflows this means that $|x| > 1$, so that $|ax^2|$ is larger than $|ax|$. In both cases, we see that an intermediate underflow or overflow is possible only when $|ax^2|$ is below the underflow threshold or over the overflow threshold: a spurious underflow or overflow is therefore impossible. This is obviously not the case when we use the scheme $a \cdot (x \cdot x)$: a simple example in binary64 arithmetic is $x = 2^{600}$ and $a = 2^{-500}$.

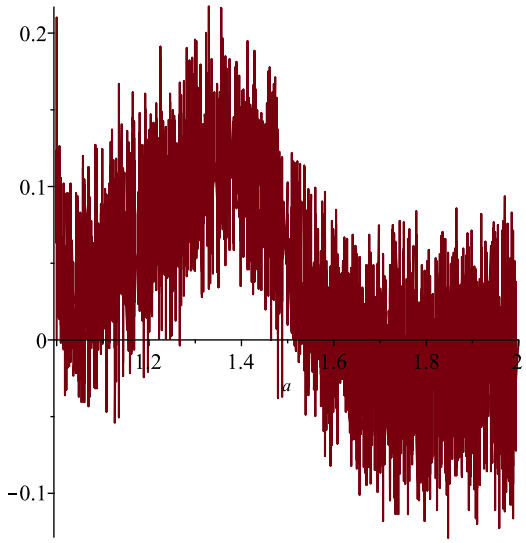

Fig. 4: Actual relative error (multiplied by $1/u$) of $(a \cdot x) \cdot x$ minus actual relative error of $a \cdot (x \cdot x)$ for $a \in [1, 2)$ in the case $p = 16$.

Hence what we should recommend to programmers is:

- for applications where avoiding spurious underflows or overflows is most important, use the scheme $(a \cdot x) \cdot x$;
- for applications where accuracy is most important, use the scheme $a \cdot (x \cdot x)$ (or, maybe better, use the scheme $a \cdot (x \cdot x)$ for $\mu(a) < 2^{2/3} \approx 1.587$ and the scheme $(a \cdot x) \cdot x$ for $\mu(a) > 2^{2/3}$: at least in the second case we are protected from the risk of spurious underflows or overflows).

Example: consider calculation of function $x \to 3x^2$ (i.e., $\mu(a) = 1.5$). If we use the scheme $a \cdot (x \cdot x)$, the relative error bound (given by (3)) is $1.75u$, and an exhaustive test in binary32 arithmetic shows that (unless underflow/overflow occurs) the largest attained error is $\approx 1.74826 \cdot u$. If we use the scheme $(a \cdot x) \cdot x$, the relative error bound given by (5) is around $1.816u$, and the largest attained error in binary32 arithmetic is $\approx 1.814977u$.

II. Saying something about axy?

What we have just done partly generalizes to the computation of axy. Using a reasoning very similar to the one of Section I-A, one easily shows that if we evaluate axy as $a \cdot (x \cdot y)$, the relative error bound (3) still applies. On the other hand, if we evaluate axy as $(a \cdot x) \cdot y$ one easily builds values x and y so that $\mu(ax)$ and $\mu(axy)$ are very close to 1 so that the only relative error bound one can deduce is $2u$ (i.e., independent of a). So what we can recommend is: if $\mu(a)$ is far enough from 1 and 2 (i.e., we are somewhere in the middle of the graph of Figure 1—a typical example is the calculation of $3xy$) it is worth using expression $a \cdot (x \cdot y)$, otherwise, the choice does not matter. Unless for the application being considered we have some a priori information on the order of magnitude of x and y, we cannot say anything on spurious underflow or overflow.

III. Computation of ax^3

Let us now consider the computation of ax^3 where a is a constant. Again, for symmetry reasons, we assume that a and x are positive. We also assume that they are (strictly) between 1 and 2. This is done without loss of generality provided that no underflow or overflow occurs in the real calculation. Different parenthesizings can be considered: $(a \cdot x) \cdot (x \cdot x)$, $((a \cdot x) \cdot x) \cdot x)$, $(a \cdot (x \cdot x)) \cdot x)$, and $a \cdot (x \cdot (x \cdot x))$. Here, we will focus on the first two, because they have properties that may make them preferable for numerical programers: the first one favors parallelism, and the second one allows one to avoid possible spurious underflows and overflows (for the same reason as the one presented in Section I-C).

A. First choice: $(a \cdot x) \cdot (x \cdot x)$

What is actually computed is $\mathrm{RN}\left(\mathrm{RN}(ax) \cdot \mathrm{RN}(x^2)\right)$. We have already considered the relative error bounds for the evaluation of $\pi_1 = x^2$ (see (1)), and $\pi_2 = ax$ (see (4)). The relative error of the last operation $\pi_1 \cdot \pi_2$ is bounded by $u/\mu(\pi_1\pi_2)$. Since the product $\pi_1\pi_2$ is between 1 and 16, this gives the following relative error bound for the last operation:

$$\begin{cases} \frac{u}{ax^3} & \text{if} \quad x < \left(\frac{2}{a}\right)^{1/3} \\ \frac{2u}{ax^3} & \text{if} \quad \left(\frac{2}{a}\right)^{1/3} < x < \left(\frac{4}{a}\right)^{1/3} \\ \frac{4u}{ax^3} & \text{if} \quad \left(\frac{4}{a}\right)^{1/3} < x < \frac{2}{a^{1/3}} \\ \frac{8u}{ax^3} & \text{if} \quad x > \frac{2}{a^{1/3}}. \end{cases} \tag{6}$$

To split the interval $[1, 2)$ (for a) into subintervals where the various bounds of (1), (4), and (6) are constant, we need to order the comparison constants of these equations, i.e., the numbers $\sqrt{2}$, $\frac{2}{a}$, $\left(\frac{2}{a}\right)^{1/3}$, $\left(\frac{4}{a}\right)^{1/3}$, and $\frac{2}{a^{1/3}}$. One easily shows (and this is illustrated by Figure 5) that:

- if $a < \sqrt{2}$ then
$$\left(\frac{2}{a}\right)^{1/3} < \sqrt{2} < \left(\frac{4}{a}\right)^{1/3} < \frac{2}{a} < \frac{2}{a^{1/3}};$$
- if $a > \sqrt{2}$ then
$$\left(\frac{2}{a}\right)^{1/3} < \frac{2}{a} < \left(\frac{4}{a}\right)^{1/3} < \sqrt{2} < \frac{2}{a^{1/3}}.$$

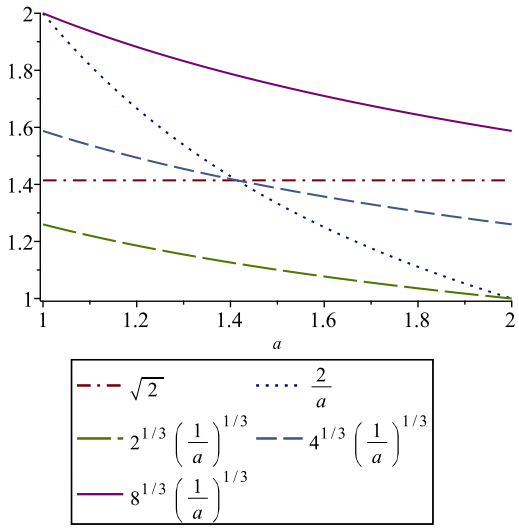

Fig. 5: The ordering of the comparison constants of (1), (4), and (6) changes at $a = \sqrt{2}$.

We can now split the interval $[1, 2)$ where x lies into 6 subintervals.

- if $a < \sqrt{2}$ then the sum of the relative errors and its maximal value in each subinterval are given in Table III.
One deduces that the relative error bound in that case is
$$u \times \max\left\{1 + \tfrac{2}{a}; 1 + \left(\tfrac{1}{2a^2}\right)^{\frac{1}{3}} + \left(\tfrac{a}{2}\right)^{\frac{2}{3}}; 1 + \tfrac{\sqrt{2}}{a}; 1 + \left(\tfrac{1}{4a^2}\right)^{\frac{1}{3}} + \left(\tfrac{a^2}{2}\right)^{\frac{1}{3}}; 1 + a^2; 1 + \tfrac{1}{a^{\frac{2}{3}}} + \tfrac{a^{\frac{2}{3}}}{2}\right\}$$
which is equal to
$$u \times \begin{cases} 1 + \frac{2}{a} & \text{if} \quad a < 2^{1/3} \approx 1.25992; \\ 1 + a^2 & \text{if} \quad a > 2^{1/3}. \end{cases} \tag{7}$$
- if $a > \sqrt{2}$ then the sum of the relative errors and its maximal value in each subinterval are given in Table IV.
One deduces that the relative error bound in that case is
$$u \times \max\left\{1 + \tfrac{2}{a}; 1 + \left(\tfrac{1}{2a^2}\right)^{\frac{1}{3}} + \left(\tfrac{a}{2}\right)^{\frac{2}{3}}; 1 + \tfrac{a^2}{2}; 1 + \left(\tfrac{a}{4}\right)^{\frac{2}{3}} + \left(\tfrac{2}{a^2}\right)^{\frac{1}{3}}; 1 + \tfrac{2\sqrt{2}}{a}; 1 + \tfrac{1}{a^{\frac{2}{3}}} + \tfrac{a^{\frac{2}{3}}}{2}\right\}$$

which is equal to

$$u \times \begin{cases} 1+\frac{2\sqrt{2}}{a} & \text{if} \quad a < 2^{5/6} \approx 1.7818; \\ 1+\frac{a^2}{2} & \text{if} \quad a > 2^{5/6}. \end{cases} \tag{8}$$

Hence, we conclude from (7) and (8) that the relative error bound of the calculation of ax^3 as $(a \cdot x) \cdot (x \cdot x)$ is

$$u \times \begin{cases} 1+\frac{2}{a} & \text{if} \quad a < 2^{1/3} \approx 1.25992; \\ 1+a^2 & \text{if} \quad 2^{1/3} < a < \sqrt{2}; \\ 1+\frac{2\sqrt{2}}{a} & \text{if} \quad \sqrt{2} < a < 2^{5/6} \approx 1.7818; \\ 1+\frac{a^2}{2} & \text{if} \quad 2^{5/6} < a. \end{cases} \tag{9}$$

Figure 6 plots the bound (9) and the actually obtained largest relative error (for all x) in the case $p = 16$ for $a \in [1, 2)$. The plot shows that the bound (9) is tight.

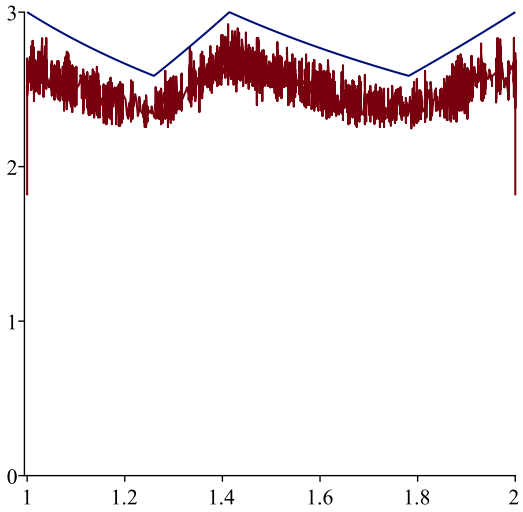

Fig. 6: The error bound for the computation of ax^3 as $(a \cdot x) \cdot (x \cdot x)$, in multiples of u and as a function of $a \in [1, 2)$, along with the largest actually obtained values for $p = 16$.

B. Second choice: $((a \cdot x) \cdot x) \cdot x)$

We start by computing $(a \cdot x) \cdot x$: the relative error of that calculation was studied in Section I-B, with bounds given for the various subintervals in Table II. The relative error of the last operation is bounded by $u/\mu(ax^3)$, i.e., it is given by (6). As previously, we need to order the comparison constants of Table II and (6).

- if $a < \sqrt{2}$ then

$$\left(\frac{2}{a}\right)^{1/3} < \left(\frac{2}{a}\right)^{1/2} < \left(\frac{4}{a}\right)^{1/3} < \frac{2}{a} < \frac{2}{\sqrt{a}} < \frac{2}{a^{1/3}},$$

and the calculation of the error bounds in the subintervals is depicted in Table V.

We deduce that the relative error is bounded by

$$u \times \max\left\{\frac{3}{a}; \frac{2a^{2/3}+2^{1/3}a^{1/3}+2^{2/3}}{2a^{2/3}}; \frac{a\sqrt{2}+\sqrt{2}+2\sqrt{a}}{2\sqrt{a}}; \frac{2a^{2/3}+2^{2/3}a^{1/3}+2^{1/3}}{2a^{2/3}}; 1+\frac{a}{2}+\frac{a^2}{2}; \frac{1}{\sqrt{a}}+1+\frac{\sqrt{a}}{2}; 1+\frac{1}{a^{1/3}}+\frac{1}{a^{2/3}}\right\},$$

which is equal to

$$u \times \begin{cases} 1+\frac{1}{a^{1/3}}+\frac{1}{a^{2/3}} & \text{if} \quad a < 1.405198; \\ 1+\frac{a}{2}+\frac{a^2}{2} & \text{if} \quad a > 1.405198. \end{cases} \tag{10}$$

- if $a > \sqrt{2}$ then

$$\left(\frac{2}{a}\right)^{1/3} < \left(\frac{2}{a}\right)^{1/2} < \frac{2}{a} < \left(\frac{4}{a}\right)^{1/3} < \frac{2}{\sqrt{a}} < \frac{2}{a^{1/3}}.$$

and the calculation of the error bounds in the subintervals is depicted in Table VI.
We deduce that the relative error is bounded by

$$u \times \max\left\{\frac{3}{a}; \frac{2a^{2/3}+2^{1/3}a^{1/3}+2^{2/3}}{2a^{2/3}}; \frac{a\sqrt{2}+\sqrt{2}+2\sqrt{a}}{2\sqrt{a}}; 1+\frac{a}{2}+\frac{a^2}{4}; \frac{2^{2/3}a^{1/3}+2a^{2/3}+2^{4/3}}{2a^{2/3}}; \frac{1}{\sqrt{a}}+1+\frac{\sqrt{a}}{2}; 1+\frac{1}{a^{1/3}}+\frac{1}{a^{2/3}}\right\},$$

which is equal to

$$u \times \begin{cases} \frac{2^{2/3}a^{1/3}+2a^{2/3}+2^{4/3}}{2a^{2/3}} & \text{if} \quad a < 1.68744; \\ 1+\frac{a}{2}+\frac{a^2}{4} & \text{if} \quad a > 1.68744. \end{cases} \tag{11}$$

Regrouping (10) and (11), we conclude that the error bound of the calculation of ax^3 as $((a \cdot x) \cdot x) \cdot x$ is

$$u \times \begin{cases} 1+\frac{1}{a^{1/3}}+\frac{1}{a^{2/3}} & \text{if} \quad a < 1.405198; \\ 1+\frac{a}{2}+\frac{a^2}{2} & \text{if} \quad 1.405198 < a < \sqrt{2}; \\ \frac{2^{2/3}a^{1/3}+2a^{2/3}+2^{4/3}}{2a^{2/3}} & \text{if} \quad \sqrt{2} < a < 1.68744; \\ 1+\frac{a}{2}+\frac{a^2}{4} & \text{if} \quad 1.68744 < a. \end{cases} \tag{12}$$

Figure 7 plots the bound (12) and the actually obtained largest relative error (for all x) in the case $p = 16$ for $a \in [1, 2)$. The plot shows that the bound (12) is tight.

C. Comparing both strategies

The error bounds of both strategies are plotted, as functions of $a \in [1, 2]$, in Figure 8. The two curves cross at $a \approx 1.32007$ and $a \approx 1.7394$. Comparing these curves leads to the following suggestion:

- if accuracy is what matters most, use $(a \cdot x) \cdot (x \cdot x)$ if $\mu(a)$ is in $[1, 1.3)$ or $[1.7, 2)$, and $((a \cdot x) \cdot x) \cdot x$ if $\mu(a) \in [1.3, 1.7]$;

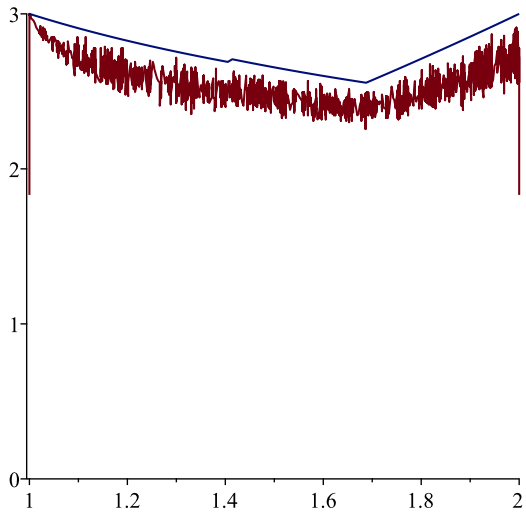

Fig. 7: The error bound for the computation of ax^3 as $((a \cdot x) \cdot x) \cdot x$, in multiples of u and as a function of $a \in [1, 2)$, along with the largest actually obtained values for $p = 16$.

- if avoiding spurious underflow or overflow is what matters most, always use $((a \cdot x) \cdot x) \cdot x$;
- if parallelism is what matters most, always use $(a \cdot x) \cdot (x \cdot x)$.

Hence, there is no "general solution" that is always the best: depending on the application, one may know that the variables lie in some restricted range (so that underflows and overflows are just impossible), or that spurious underflows and overflows must be avoided at all costs.

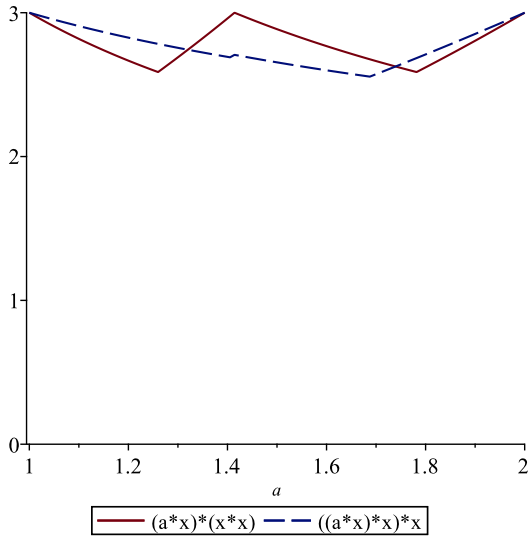

Fig. 8: The error bounds of the two schemes for ax^3 studied in this paper: $(a \cdot x) \cdot (x \cdot x)$ (plain line) and $((a \cdot x) \cdot x) \cdot x$ (dashed line). The two curves cross at $a \approx 1.32007$ and $a \approx 1.7394$.

Example: consider calculation of function $x \to 3x^3$ (i.e., $\mu(a) = 1.5$). If we evaluate the function as $(a \cdot x) \cdot (x \cdot x)$, the error bound given by (9) is $2.886u$ and the largest attained error in binary32 arithmetic is $2.865u$. If we evaluate the function as $((a \cdot x) \cdot x) \cdot x$, the error bound given by (12) is $2.655u$ and the largest attained error in binary32 arithmetic is $2.612u$.

Conclusion

We have shown through small examples that when evaluating some expressions that contain a constant, it may be interesting to use different orders of evaluation depending on the value of the constant. Such a choice may be taken by the programmer or, when the programmer does not indicate an unambiguous order through parenthesizing and the language specification does not impose an order of evaluation, by the compiler.

References

[1] IEEE. *IEEE Standard for Floating-Point Arithmetic (IEEE Std 754-2019)*. July 2019.

[2] International Organization for Standardization. *Programming languages – Fortran – Part 1: Base language*. International Standard ISO/IEC 1539-1:2004, 2004.

[3] Claude-Pierre Jeannerod and Siegfried M. Rump. On relative errors of floating-point operations: optimal bounds and applications. *Mathematics of Computation*, (87):803–819, 2018. https://hal.inria.fr/hal-00934443.

[4] D. Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley, Reading, MA, 3rd edition, 1998.

[5] Jean-Michel Muller, Nicolas Brunie, Florent de Dinechin, Claude-Pierre Jeannerod, Mioara Joldes, Vincent Lefèvre, Guillaume Melquiond, Nathalie Revol, and Serge Torres. *Handbook of Floating-Point Arithmetic, 2nd edition*. Birkhäuser Boston, 2018. ACM G.1.0; G.1.2; G.4; B.2.0; B.2.4; F.2.1., ISBN 978-3-319-76525-9.

[6] S. M. Rump. Error bounds for computer arithmetics. In *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, pages 1–14, 2019.

TABLE III: Largest relative error in each of the subintervals (we compute $(a \cdot x) \cdot (x \cdot x)$, and a is less than $\sqrt{2}$).

$x \in$	$\left[1, \left(\frac{2}{a}\right)^{1/3}\right]$	$\left[\left(\frac{2}{a}\right)^{1/3}, \sqrt{2}\right]$	$\left[\sqrt{2}, \left(\frac{4}{a}\right)^{1/3}\right]$	$\left[\left(\frac{4}{a}\right)^{1/3}, \frac{2}{a}\right]$	$\left[\frac{2}{a}, \frac{2}{a^{1/3}}\right]$	$\left[\frac{2}{a^{1/3}}, 2\right)$
$\frac{\text{relative error}}{u}$	$\frac{1}{x^2} + \frac{1}{ax} + \frac{1}{ax^3}$	$\frac{1}{x^2} + \frac{1}{ax} + \frac{2}{ax^3}$	$\frac{2}{x^2} + \frac{1}{ax} + \frac{2}{ax^3}$	$\frac{2}{x^2} + \frac{1}{ax} + \frac{4}{ax^3}$	$\frac{2}{x^2} + \frac{2}{ax} + \frac{4}{ax^3}$	$\frac{2}{x^2} + \frac{2}{ax} + \frac{8}{ax^3}$
$\frac{\text{largest value}}{u}$	$1 + \frac{2}{a}$	$1 + \left(\frac{1}{2a^2}\right)^{\frac{1}{3}} + \left(\frac{a}{2}\right)^{\frac{2}{3}}$	$1 + \frac{\sqrt{2}}{a}$	$1 + \left(\frac{1}{4a^2}\right)^{\frac{1}{3}} + \left(\frac{a^2}{2}\right)^{\frac{1}{3}}$	$1 + a^2$	$1 + \frac{1}{a^{\frac{2}{3}}} + \frac{a^{\frac{2}{3}}}{2}$

TABLE IV: Largest relative error in each of the subintervals (we compute $(a \cdot x) \cdot (x \cdot x)$, and a is larger than $\sqrt{2}$).

$x \in$	$\left[1, \left(\frac{2}{a}\right)^{1/3}\right]$	$\left[\left(\frac{2}{a}\right)^{1/3}, \frac{2}{a}\right]$	$\left[\frac{2}{a}, \left(\frac{4}{a}\right)^{1/3}\right]$	$\left[\left(\frac{4}{a}\right)^{1/3}, \sqrt{2}\right]$	$\left[\sqrt{2}, \frac{2}{a^{1/3}}\right]$	$\left[\frac{2}{a^{1/3}}, 2\right)$
$\frac{\text{relative error}}{u}$	$\frac{1}{x^2} + \frac{1}{ax} + \frac{1}{ax^3}$	$\frac{1}{x^2} + \frac{1}{ax} + \frac{2}{ax^3}$	$\frac{1}{x^2} + \frac{2}{ax} + \frac{2}{ax^3}$	$\frac{1}{x^2} + \frac{2}{ax} + \frac{4}{ax^3}$	$\frac{2}{x^2} + \frac{2}{ax} + \frac{4}{ax^3}$	$\frac{2}{x^2} + \frac{2}{ax} + \frac{8}{ax^3}$
$\frac{\text{largest value}}{u}$	$1 + \frac{2}{a}$	$1 + \left(\frac{1}{2a^2}\right)^{\frac{1}{3}} + \left(\frac{a}{2}\right)^{\frac{2}{3}}$	$1 + \frac{a^2}{2}$	$1 + \left(\frac{a}{4}\right)^{\frac{2}{3}} + \left(\frac{2}{a^2}\right)^{\frac{1}{3}}$	$1 + \frac{2\sqrt{2}}{a}$	$1 + \frac{1}{a^{\frac{2}{3}}} + \frac{a^{\frac{2}{3}}}{2}$

TABLE V: Largest relative error in each of the subintervals (we compute $((a \cdot x) \cdot x) \cdot x$, and a is less than $\sqrt{2}$).

$x \in$	$\left[1, \left(\frac{2}{a}\right)^{1/3}\right]$	$\left[\left(\frac{2}{a}\right)^{1/3}, \left(\frac{2}{a}\right)^{1/2}\right]$	$\left[\left(\frac{2}{a}\right)^{1/2}, \left(\frac{4}{a}\right)^{1/3}\right]$	$\left[\left(\frac{4}{a}\right)^{1/3}, \frac{2}{a}\right]$	$\left[\frac{2}{a}, \frac{2}{\sqrt{a}}\right]$	$\left[\frac{2}{\sqrt{a}}, \frac{2}{a^{1/3}}\right]$	$\left[\frac{2}{a^{1/3}}, 2\right)$
$\frac{\text{relative error}}{u}$	$\frac{1}{ax} + \frac{1}{ax^2} + \frac{1}{ax^3}$	$\frac{1}{ax} + \frac{1}{ax^2} + \frac{2}{ax^3}$	$\frac{1}{ax} + \frac{2}{ax^2} + \frac{2}{ax^3}$	$\frac{1}{ax} + \frac{2}{ax^2} + \frac{4}{ax^3}$	$\frac{2}{ax} + \frac{2}{ax^2} + \frac{4}{ax^3}$	$\frac{2}{ax} + \frac{4}{ax^2} + \frac{4}{ax^3}$	$\frac{2}{ax} + \frac{4}{ax^2} + \frac{8}{ax^3}$
$\frac{\text{largest value}}{u}$	$\frac{3}{a}$	$\frac{2a^{2/3}+2^{1/3}a^{1/3}+2^{2/3}}{2a^{2/3}}$	$\frac{a\sqrt{2}+\sqrt{2}+2\sqrt{a}}{2\sqrt{a}}$	$\frac{2a^{2/3}+2^{2/3}a^{1/3}+2^{1/3}}{2a^{2/3}}$	$1 + \frac{a}{2} + \frac{a^2}{2}$	$\frac{1}{\sqrt{a}} + 1 + \frac{\sqrt{a}}{2}$	$1 + \frac{1}{a^{1/3}} + \frac{1}{a^{2/3}}$

TABLE VI: Largest relative error in each of the subintervals (we compute $((a \cdot x) \cdot x) \cdot x$, and a is larger than $\sqrt{2}$).

$x \in$	$\left[1, \left(\frac{2}{a}\right)^{1/3}\right]$	$\left[\left(\frac{2}{a}\right)^{1/3}, \left(\frac{2}{a}\right)^{1/2}\right]$	$\left[\left(\frac{2}{a}\right)^{1/2}, \frac{2}{a}\right]$	$\left[\frac{2}{a}, \left(\frac{4}{a}\right)^{1/3}\right]$	$\left[\left(\frac{4}{a}\right)^{1/3}, \frac{2}{\sqrt{a}}\right]$	$\left[\frac{2}{\sqrt{a}}, \frac{2}{a^{1/3}}\right]$	$\left[\frac{2}{a^{1/3}}, 2\right)$
$\frac{\text{relative error}}{u}$	$\frac{1}{ax} + \frac{1}{ax^2} + \frac{1}{ax^3}$	$\frac{1}{ax} + \frac{1}{ax^2} + \frac{2}{ax^3}$	$\frac{1}{ax} + \frac{2}{ax^2} + \frac{2}{ax^3}$	$\frac{2}{ax} + \frac{2}{ax^2} + \frac{2}{ax^3}$	$\frac{2}{ax} + \frac{2}{ax^2} + \frac{4}{ax^3}$	$\frac{2}{ax} + \frac{4}{ax^2} + \frac{4}{ax^3}$	$\frac{2}{ax} + \frac{4}{ax^2} + \frac{8}{ax^3}$
$\frac{\text{largest value}}{u}$	$\frac{3}{a}$	$\frac{2a^{2/3}+2^{1/3}a^{1/3}+2^{2/3}}{2a^{2/3}}$	$\frac{a\sqrt{2}+\sqrt{2}+2\sqrt{a}}{2\sqrt{a}}$	$1 + \frac{a}{2} + \frac{a^2}{4}$	$\frac{2^{2/3}a^{1/3}+2a^{2/3}+2^{4/3}}{2a^{2/3}}$	$\frac{1}{\sqrt{a}} + 1 + \frac{\sqrt{a}}{2}$	$1 + \frac{1}{a^{1/3}} + \frac{1}{a^{2/3}}$